10

Y(Income)	ID	M	E	PE	H	A	G	IQ
65532	1	A	C	C	B	B	B	110
86574	2	A	A	B	A	B	C	116
32568	3	C	A	B	C	C	C	95
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
45875	N	B	B	A	B	A	C	98

Fig. 1

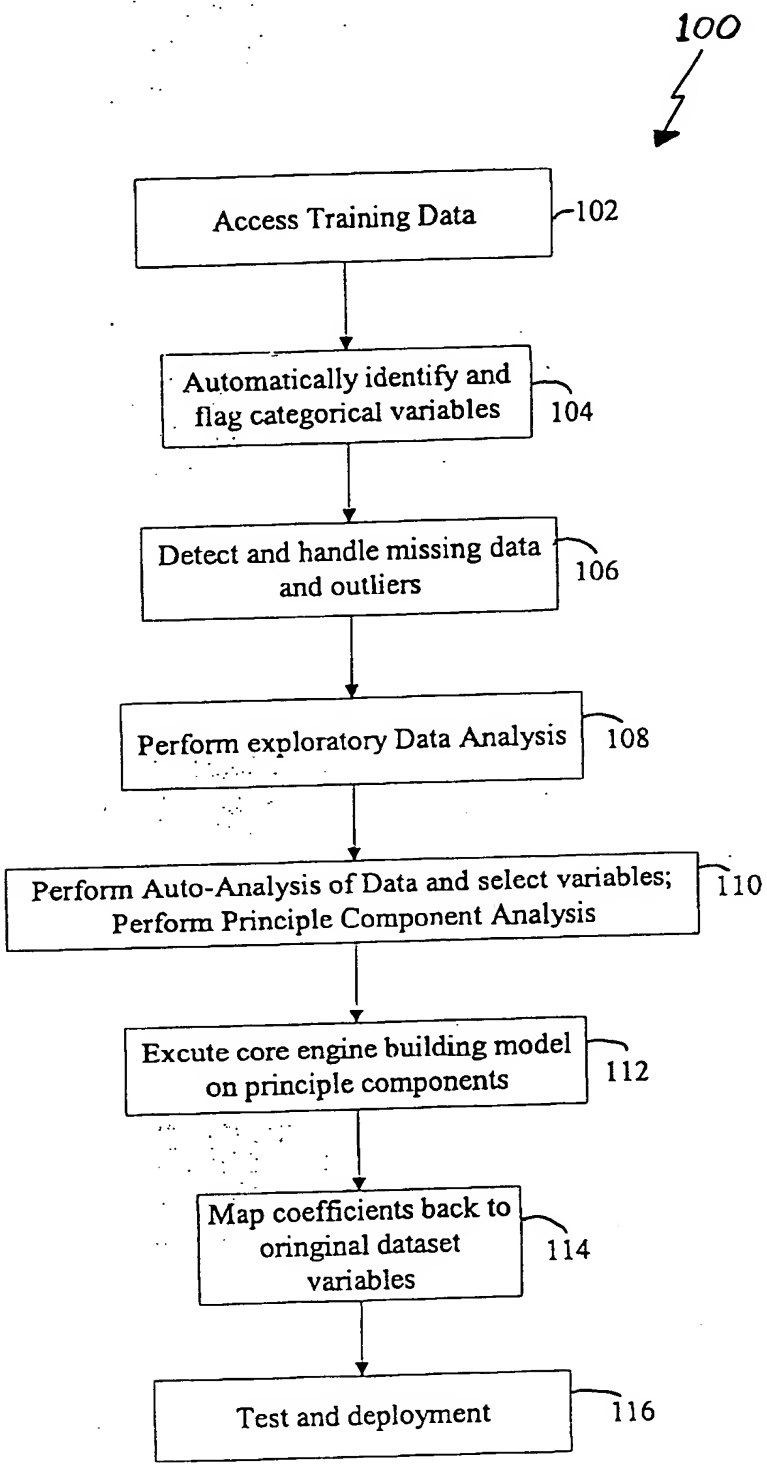
100
Access Training Data
102
Automatically identify and flag categorical variables
104
Detect and handle missing data and outliers
106
Perform exploratory Data Analysis
108
Perform Auto-Analysis of Data and select variables; Perform Principle Component Analysis
110
Excute core engine building model on principle components
112
Map coefficients back to oringinal dataset variables
114
Test and deployment
116

Fig. 2

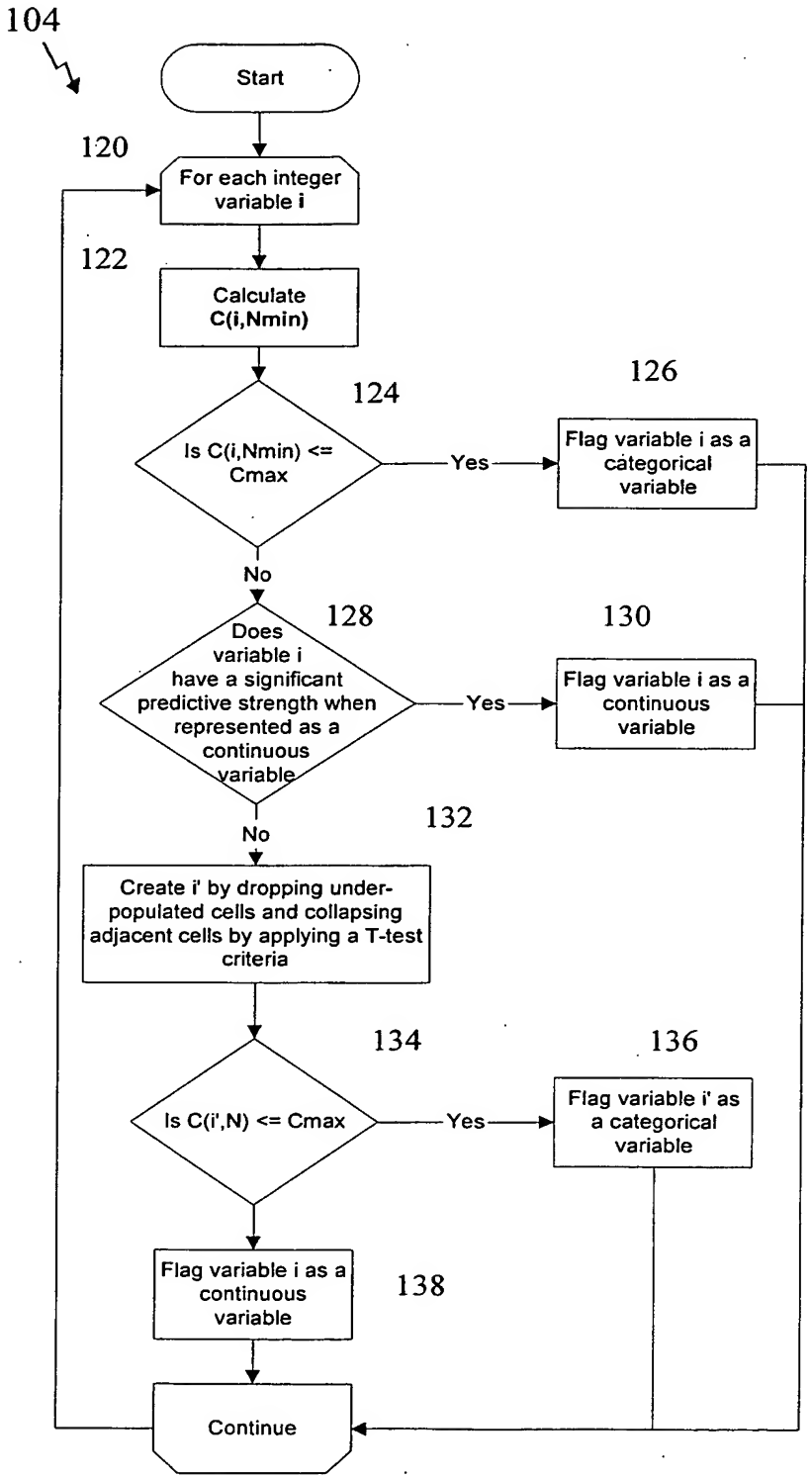
104
Start
120
For each integer variable i
122
Calculate C(i,Nmin)
124
126
Is C(i,Nmin) <= Cmax
Yes
Flag variable i as a categorical variable
No
128
130
Does variable i have a significant predictive strength when represented as a continuous variable
Yes
Flag variable i as a continuous variable
No
132
Create i' by dropping under-populated cells and collapsing adjacent cells by applying a T-test criteria
134
136
Is C(i',N) <= Cmax
Yes
Flag variable i' as a categorical variable
Flag variable i as a continuous variable
138
Continue

FIG. 3

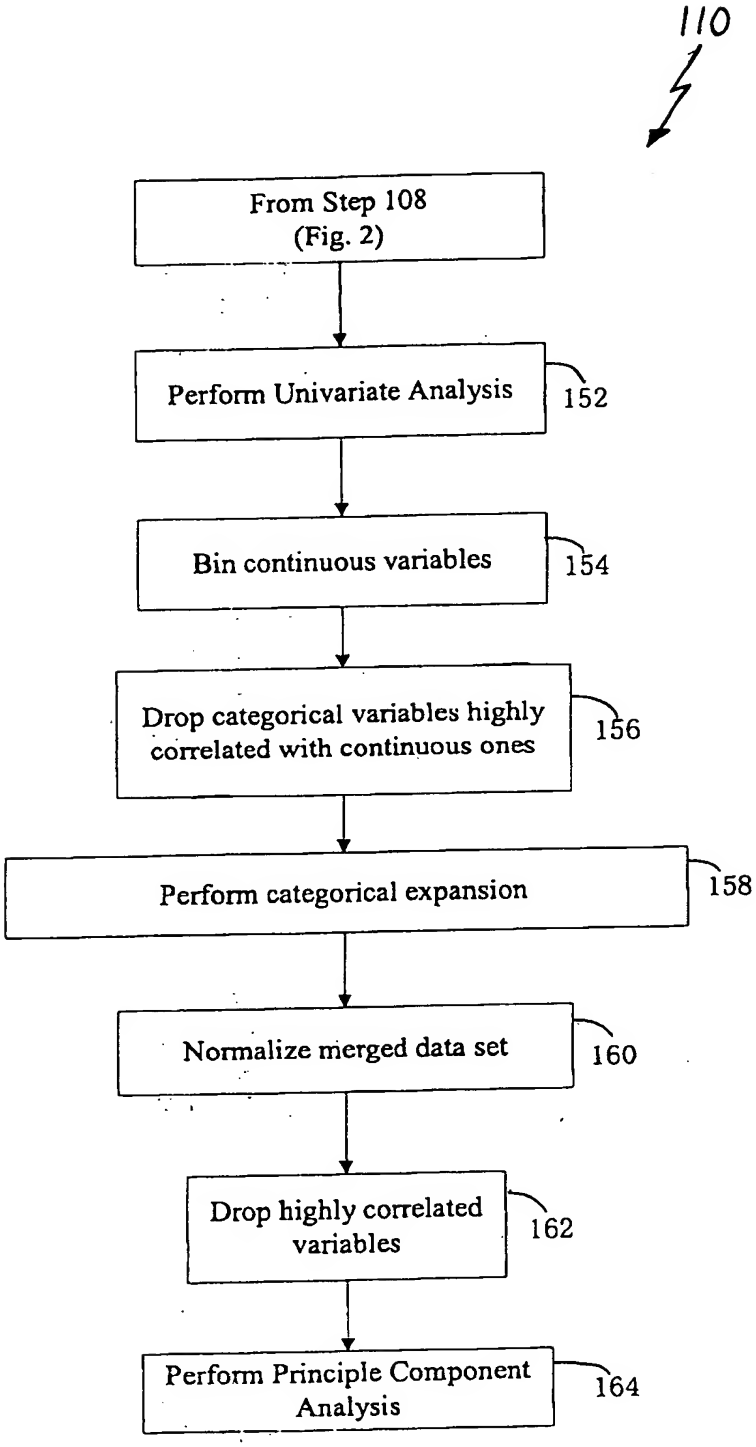
110
From Step 108 (Fig. 2)
Perform Univariate Analysis
152
Bin continuous variables
154
Drop categorical variables highly correlated with continuous ones
156
Perform categorical expansion
158
Normalize merged data set
160
Drop highly correlated variables
162
Perform Principle Component Analysis
164

Fig. 4

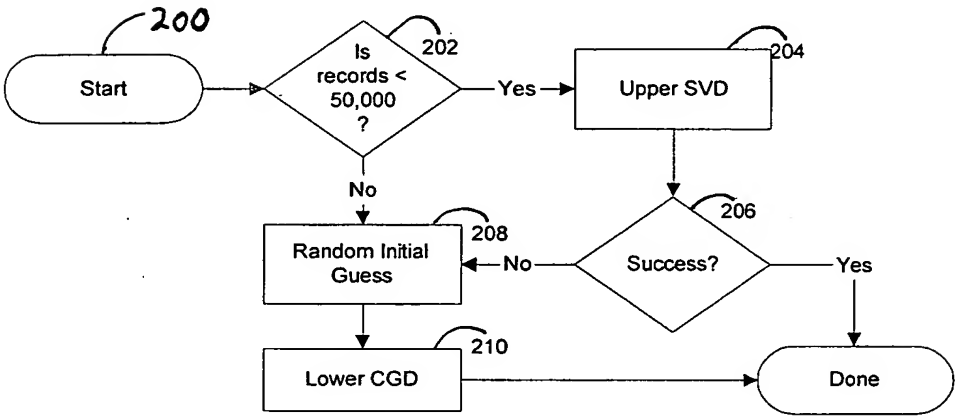
200
Start
Is records < 50,000 ?
202
Yes
Upper SVD
204
No
206
Success?
No
Yes
Random Initial Guess
208
Lower CGD
210
Done

FIG. 5

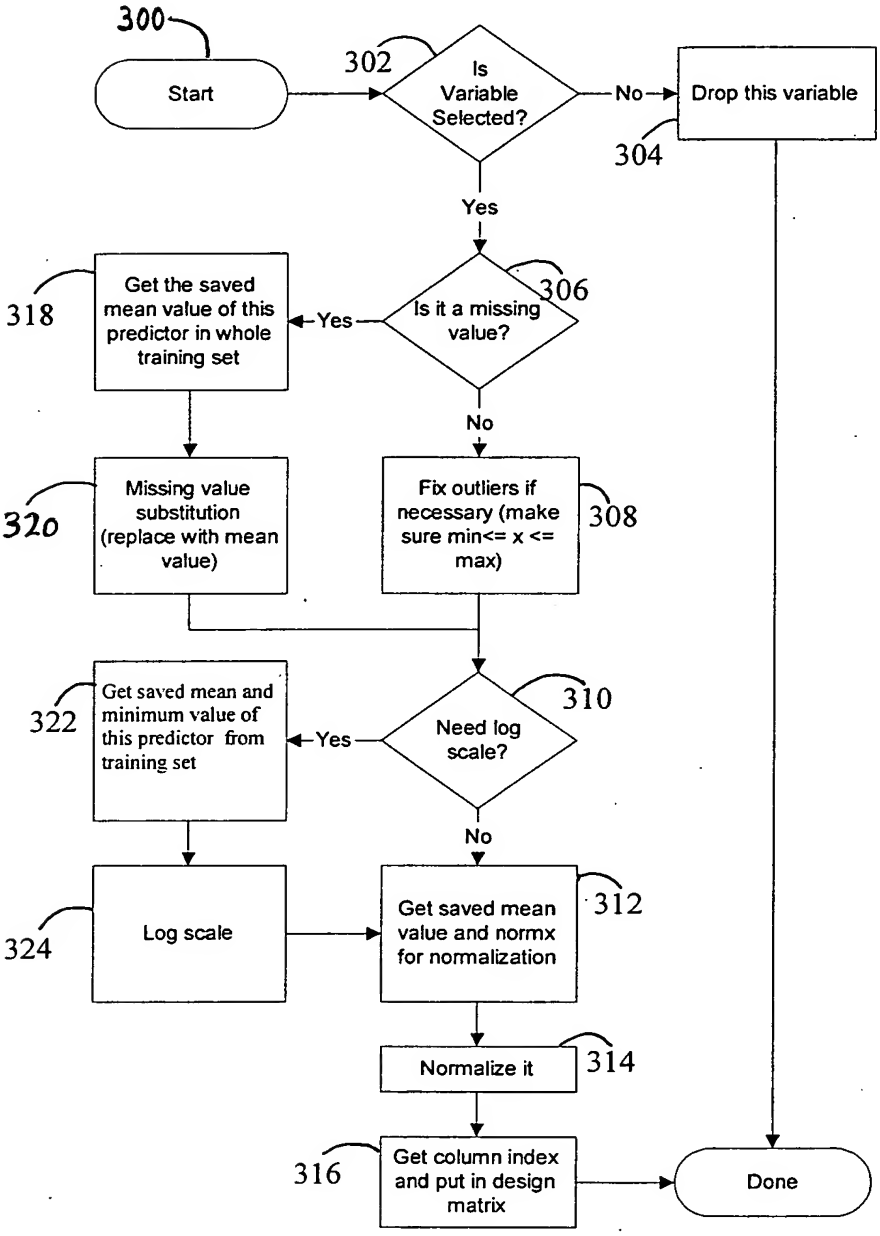
300
Start
302
Is Variable Selected?
No
Drop this variable
304
Yes
306
Is it a missing value?
Yes
318
Get the saved mean value of this predictor in whole training set
No
320
Missing value substitution (replace with mean value)
Fix outliers if necessary (make sure min<= x <= max)
308
310
Need log scale?
Yes
322
Get saved mean and minimum value of this predictor from training set
No
324
Log scale
Get saved mean value and normx for normalization
312
Normalize it
314
316
Get column index and put in design matrix
Done

FIG. 6

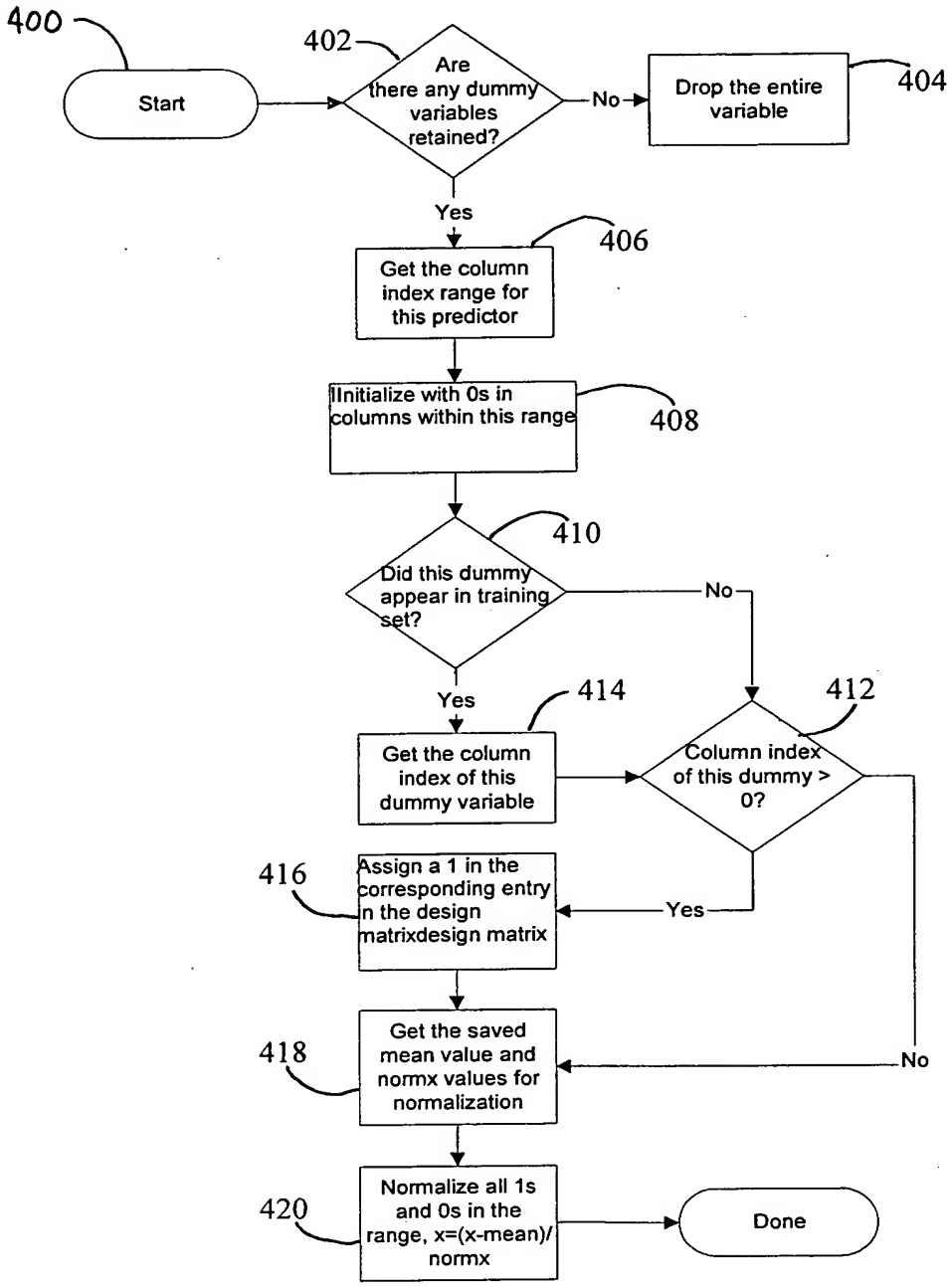
400
Start
402
Are there any dummy variables retained?
No
Drop the entire variable
404
Yes
406
Get the column index range for this predictor
Initialize with 0s in columns within this range
408
410
Did this dummy appear in training set?
No
Yes
414
412
Get the column index of this dummy variable
Column index of this dummy > 0?
416
Assign a 1 in the corresponding entry in the design matrixdesign matrix
Yes
Get the saved mean value and normx values for normalization
418
No
Normalize all 1s and 0s in the range, x=(x-mean)/ normx
420
Done

FIG. 7